Exciting promises and potential pitfalls of big data in biology and medicine (大数据于生物学和医学的前途与隐患)

Limsoon Wong (黄任祥)

何为大数据 (Big data)

- 大数据的特点
  - 数量 (volume)
  - 速度 (velocity)
  - 品种 (variety)
- 其他特点
  - 准确性 (veracity), v...

“大/多/繁 到令你无从处理”

“More than you know how to handle”

- 大数据的挑战
  - 很多规模问题 (scaling issues)
  - 但也有些影响当前的生物信息学和统计学分析的基本假设的问题 (break analysis procedures in fundamental ways that are not related to scaling issues)

提纲

- **被遗忘的假设**
  - 独立相同分布中的相同分布 (2nd “I” in I.I.D.)
- **更多不一定是更好**
  - 蛋白质复合物 (protein complexes)
  - 致病基因 (causal genes)

Forgotten assumptions

THE 2ND "I" IN I.I.D.

独立相同分布 中的相同分布

假設檢定 (Hypothesis testing)

- 假設檢定大都假设样本是根据独立相同分布(IID)绘制的

- **Commonly used statistical tests (T-test, $\chi 2$ test, Wilcoxon rank-sum test, …) all assume samples are drawn from independent identical distributions (I.I.D.)**

确保IID

- 在临床试验中，我们精心选择样品以确保IID
  - 独立分布Independent: Patients are not related
  - 相同分布Identical: Similar # of male/female, young/old, … in cases and controls

	A	B
lived	60	65
died	100	165

Thus sex, age, … don't need to appear in the contingency table

- 在大数据的分析，并在许多数据挖掘作品中，人们几乎都忽视了此点！

是也？非也？

Overall

	A	B
lived	60	65
died	100	165

A更好?
Looks like treatment A is better

Women

	A	B
lived	40	15
died	20	5

Men

	A	B
lived	20	50
died	80	160

B更好?
Looks like treatment B is better

History of heart disease

	A	B
lived	10	5
died	70	50

No history of heart disease

	A	B
lived	10	45
died	10	110

A更好?
Looks like treatment A is better

非相同分布

Overall

	A	B
lived	60	65
died	100	165

Women

	A	B
lived	40	15
died	20	5

Men

	A	B
lived	20	50
died	80	160

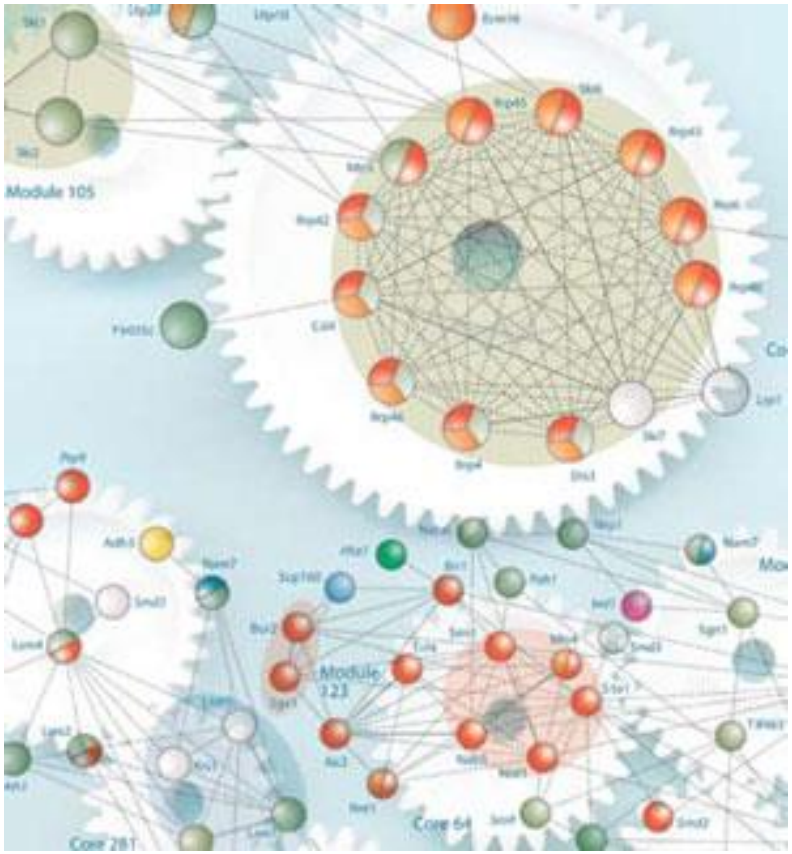
History of heart disease

	A	B
lived	10	5
died	70	50

No history of heart disease

	A	B
lived	10	45
died	10	110

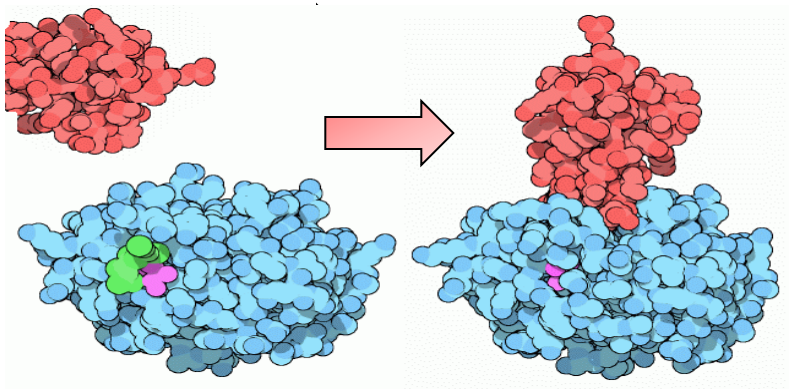
- **使用A**
  - 男 = 100 (63%)
  - 女 = 60 (37%)
- **使用B**
  - 男= 210 (91%)
  - 女= 20 (9%)
- **男使用A**
  - 有病历 = 80 (80%)
  - 无病历 = 20 (20%)
- **男使用B**
  - 有病历= 55 (26%)
  - 无病历= 155 (74%)



More may not be better

PROTEIN COMPLEXES
蛋白质复合物

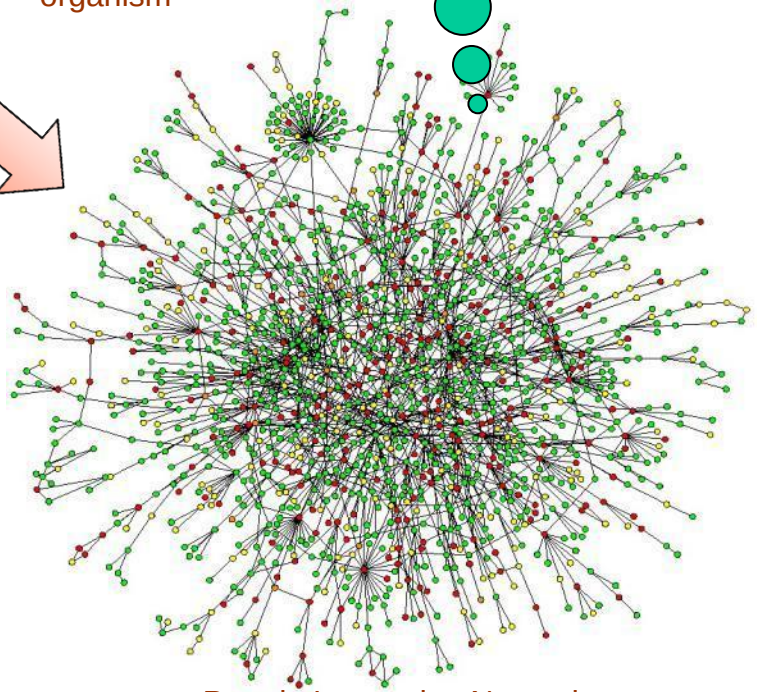
蛋白相互作用网络(PPIN)



Individual proteins come together and interact

时空信息损失了。
如何在PPIN里找
回蛋白质复合物

Collection of such interactions in an organism



Protein Interaction Network

- **Proteins come together & interact**
- **The collection of these interactions form a Protein Interaction Network or PPIN**

难处

- **Cytochrome BC1 complex**
  - Involved in electron-transport chain in mitochondrial inner membrane

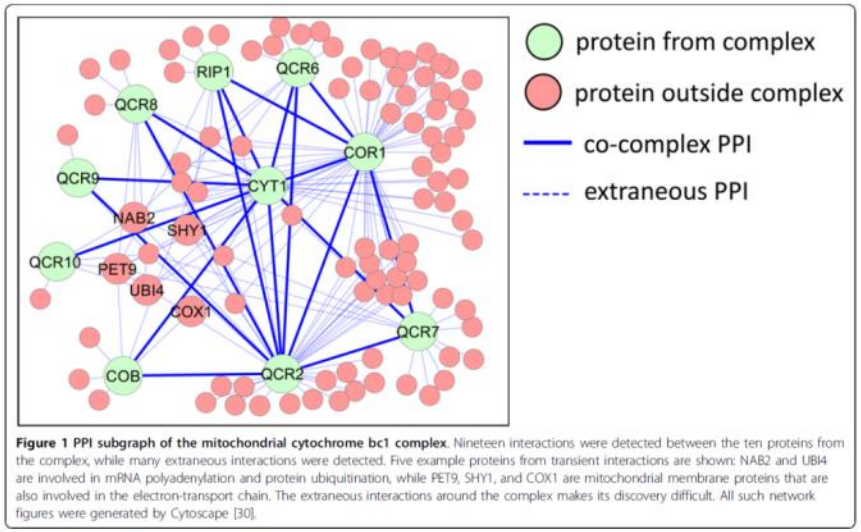


Figure 1 PPI subgraph of the mitochondrial cytochrome bc1 complex. Nineteen interactions were detected between the ten proteins from the complex, while many extraneous interactions were detected. Five example proteins from transient interactions are shown: NAB2 and UBI4 are involved in mRNA polyadenylation and protein ubiquitination, while PET9, SHY1, and COX1 are mitochondrial membrane proteins that are also involved in the electron-transport chain. The extraneous interactions around the complex makes its discovery difficult. All such network figures were generated by Cytoscape [30].

- 从蛋白相互作用网络中挖出**BC1 是非常困难的**
  - BC1在蛋白相互作用网络的子网非常稀疏
    - **BC1的蛋白质之间的45个有可能出现的相互作用，只有19个被测出**
  - 与BC1之外的太多其他蛋白质有相互作用
    - **E.g., UBI4 is involved in protein ubiquitination, and binds to many proteins to perform its function**

Yong, et al. **Supervised maximum-likelihood weighting of composite protein networks for complex prediction**. *BMC Systems Biology*, 6(Suppl 2):S13, 2012

NUS National University of Singapore

大数据能否帮上忙

- **复合网 (Composite network)**
  - 如果蛋白质u和v按照任何数据源是相关的，把u和v连上

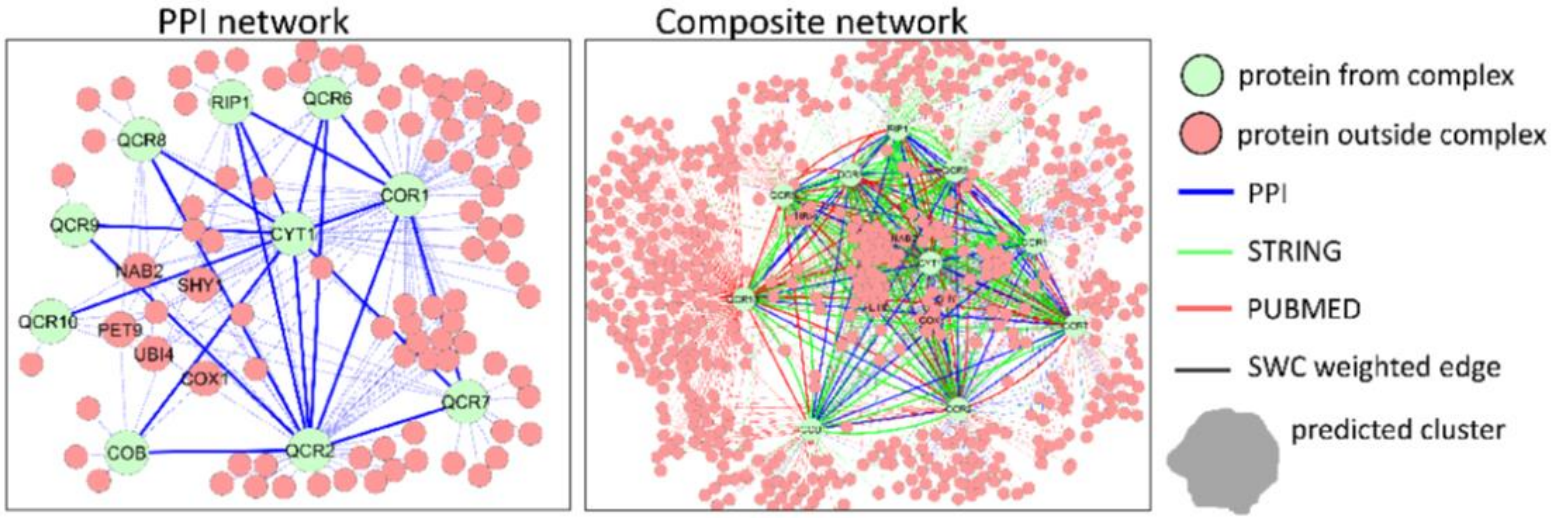
Data source	Database	Scoring method
PPI	BioGRID, IntACT, MINT	Iterative AdjustCD.
L2-PPI (indirect PPI)	BioGRID, IntACT, MINT	Iterative AdjustCD
Functional association	STRING	STRING
Literature co-occurrence	PubMed	Jaccard coefficient

	Yeast			Human		
	# Pairs	% co-complex	coverage	# Pairs	% co-complex	coverage
PPI	106328	**5.8%**	**55%**	48098	10%	14%
L2-PPI	181175	1.1%	18%	131705	5.5%	20%
STRING	175712	5.7%	89%	311435	3.1%	27%
PubMed	161213	4.9%	70%	91751	4.3%	11%
All	531800	**2.1%**	**98%**	522668	3.4%	49%

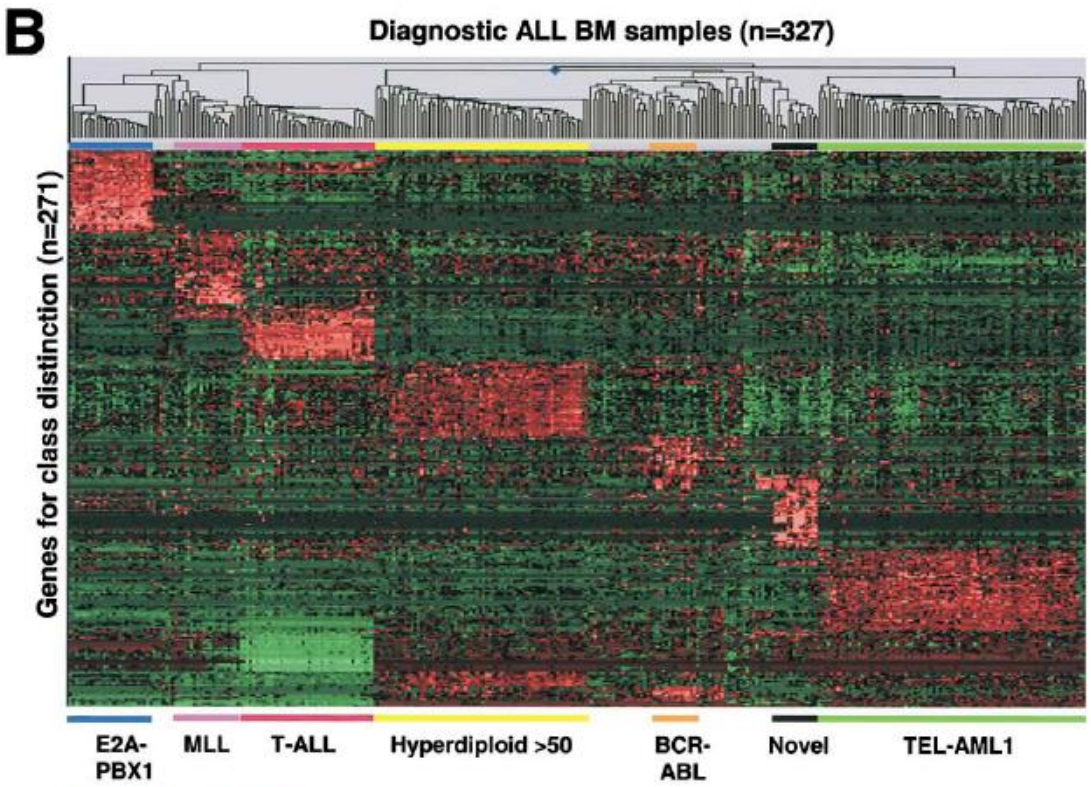
Yong, et al. **Supervised maximum-likelihood weighting of composite protein networks for complex prediction**. *BMC Systems Biology*, 6(Suppl 2):S13, 2012

NUS National University of Singapore

更多不一定是更好



BC1里的蛋白质在符合网络中完全连接，但也有太多外来蛋白质。所以除非你知道如何删除多余的蛋白质，聚类(clustering)是不会发现BC1的



More may not be better

CAUSAL GENES

致病基因

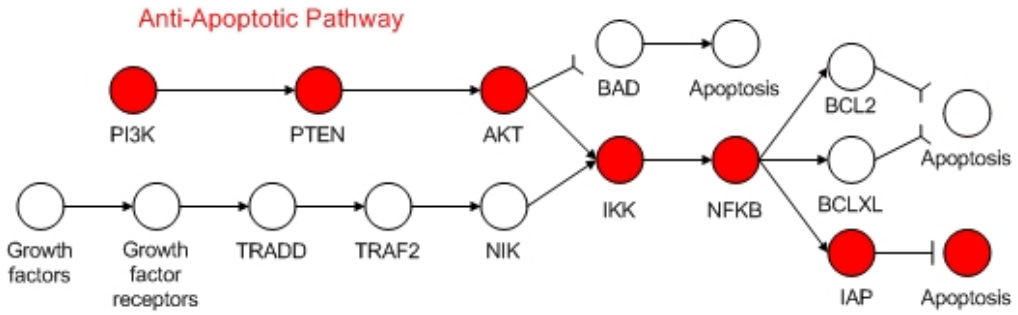
基因表达分析的挑战

- 在不同实验中有差异表达的基因的交叠百分比非常低
  - Prostate cancer
    - **Lapointe et al, 2004**
    - **Singh et al, 2002**
  - Lung cancer
    - **Garber et al, 2001**
    - **Bhattacharjee et al, 2001**
  - DMD
    - **Haslett et al, 2002**
    - **Pescatori et al, 2007**

Datasets	DEG	POG
Prostate Cancer		
	Top 10	0.30
	Top 50	0.14
	Top100	0.15
Lung Cancer		
	Top 10	0.00
	Top 50	0.20
	Top100	0.31
DMD		
	Top 10	0.20
	Top 50	0.42
	Top100	0.54

Zhang et al, *Bioinformatics*, 2009

从生物学入手



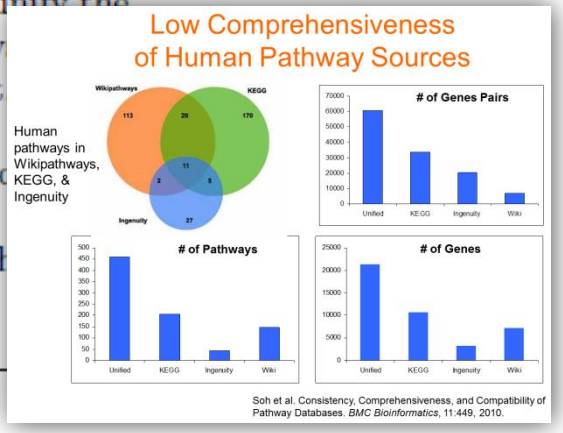
- 每种疾病都有一定的潜在原因

⇒ 真正与疾病有关的基因应该遵循一个统一的生物主题

- 选定基因的不确定度可以通过考虑的基因的生物过程（**biological processes)** 被减少

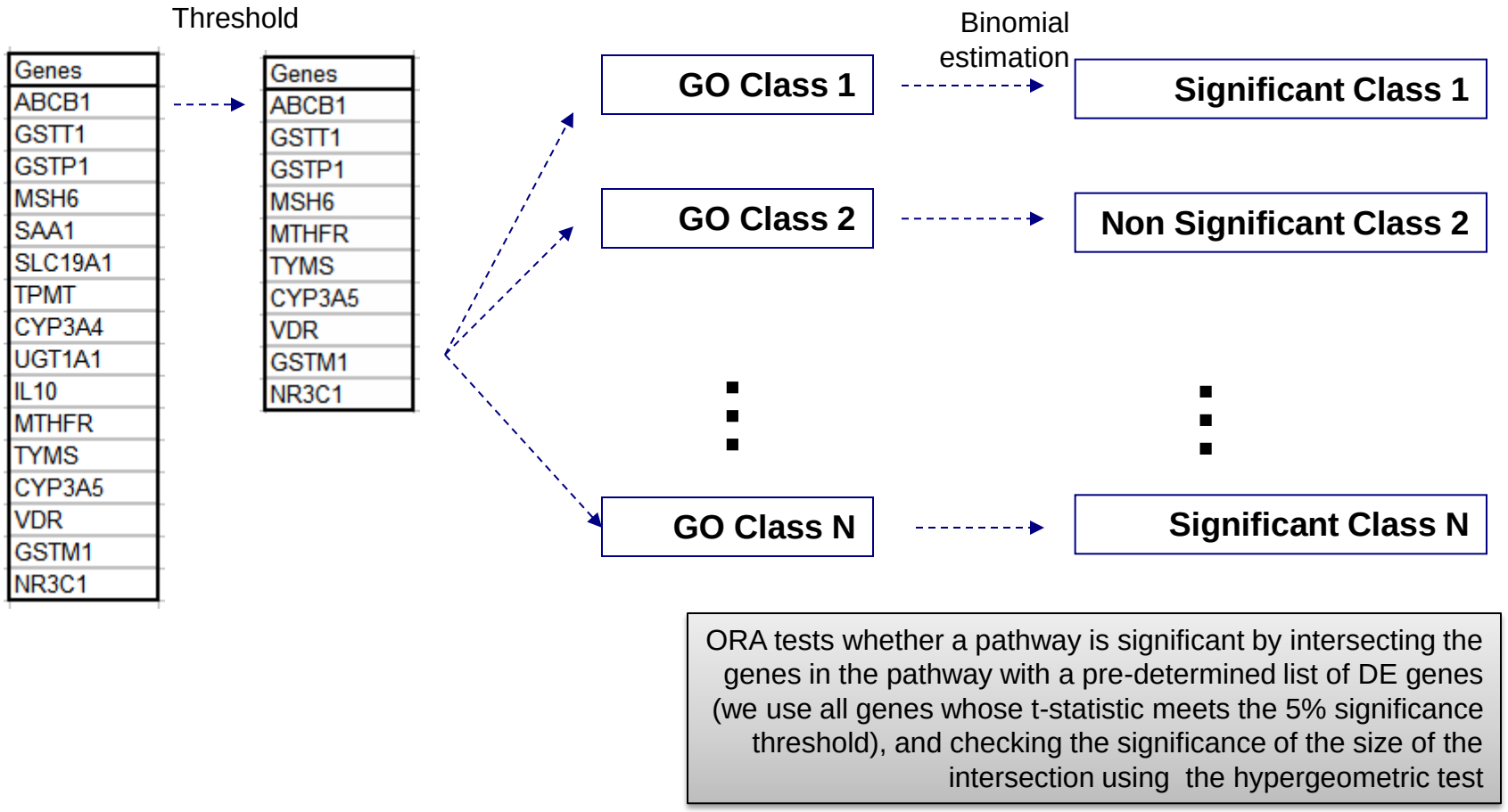
Database	Remarks
KEGG	KEGG (`http://www.genome.jp/kegg`) is one of the best known pathway databases (Kanehisa *et al.*, 2010). It consists of 16 main databases, comprising different levels of biological information such as systems, genomic, etc. The data files are downloadable in XML format. At time of writing it has 392 pathways.
WikiPathways	WikiPathways (`http://www.wikipathways.org`) is a Wikipedia-based collaborative effort among various labs (Kelder *et al.*, 2009). It has 1,627 pathways of which 369 are human. The content is downloadable in GPML format.
Reactome	Reactome (`http::/www.reactome.org`) is also a collaborative effort like WikiPathways (Vastrik *et al.*, 2007). It is one of the largest datasets, with over 4,166 human reactions organized into 1,131 pathways by December 2010. Reactome can be downloaded in BioPax and SBML among other formats.
Pathway Commons	Pathway Commons (`http://www.pathwaycommons.com`) collects information from various databases but does not unify the data (Cerami *et al.*, 2006). It contains 1,573 pathway 564 organisms. The data is returned in BioPax format
PathwayAPI	PathwayAPI (`http://www.pathwayapi.com`) contains unified human pathways obtained from a merge of WikiPathways and Ingenuity® Knowledge Base (Soh 2010). Data is downloadable as a SQL dump or as a and is also interfaceable in JSON format.

生物途径 (biological pathways) 的大数据



Goh, et al. *Proteomics*, 12(4-5):550-563, 2012.

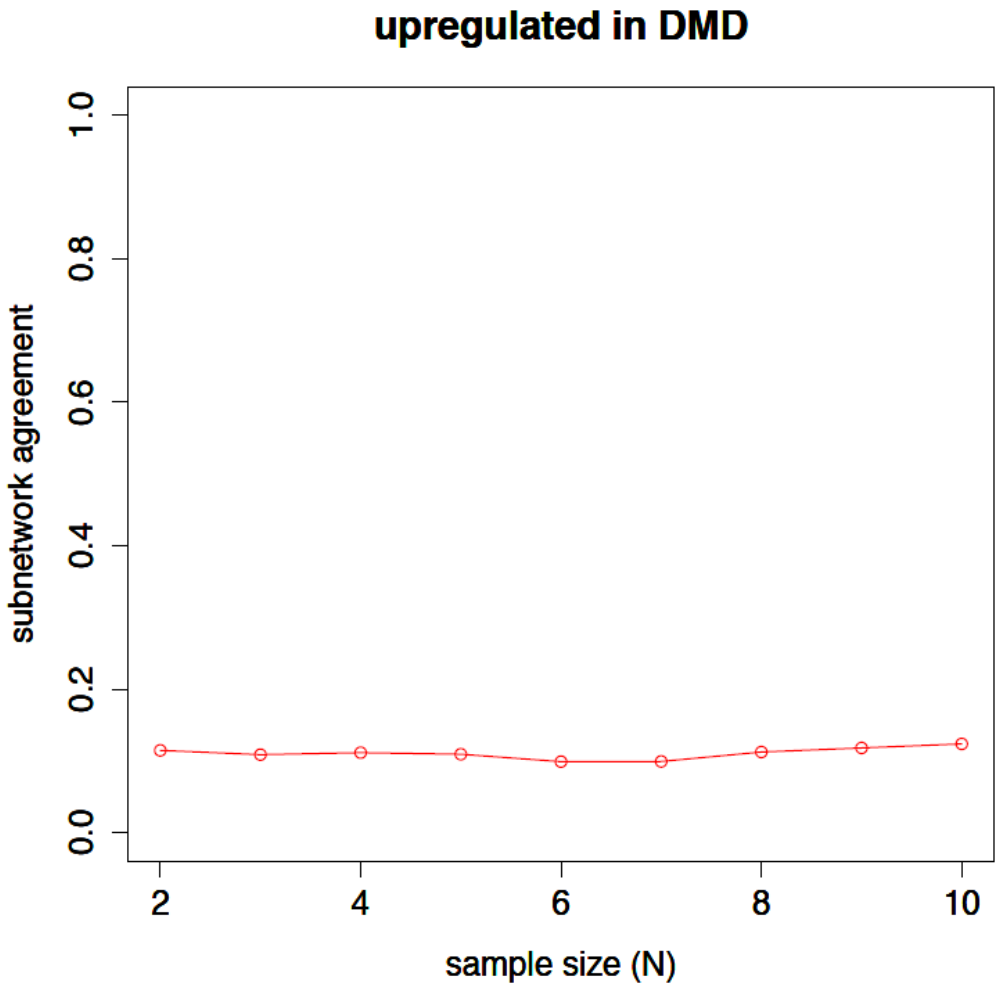
富集分析 (ORA)



ORA tests whether a pathway is significant by intersecting the genes in the pathway with a pre-determined list of DE genes (we use all genes whose t-statistic meets the 5% significance threshold), and checking the significance of the size of the intersection using the hypergeometric test

S Draghici et al. “Global functional profiling of gene expression”. *Genomics*, 81(2):98-104, 2003.

令人失望的表现



DMD gene expression data
- Pescatori et al., 2007
- Haslett et al., 2002

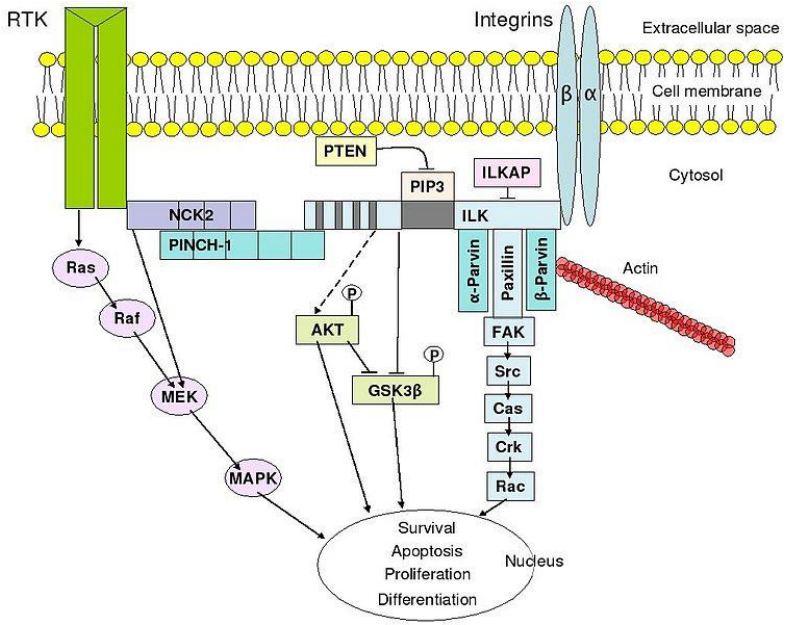
Pathway data
- PathwayAPI, Soh et al., 2010

Issue #1 with ORA

- 其虚假设(**null hypothesis**)基本上说“在给定的生物途径中的基因的行为和随机选择的基因组没有任何不同”

- 这个虚假设显然是错误的

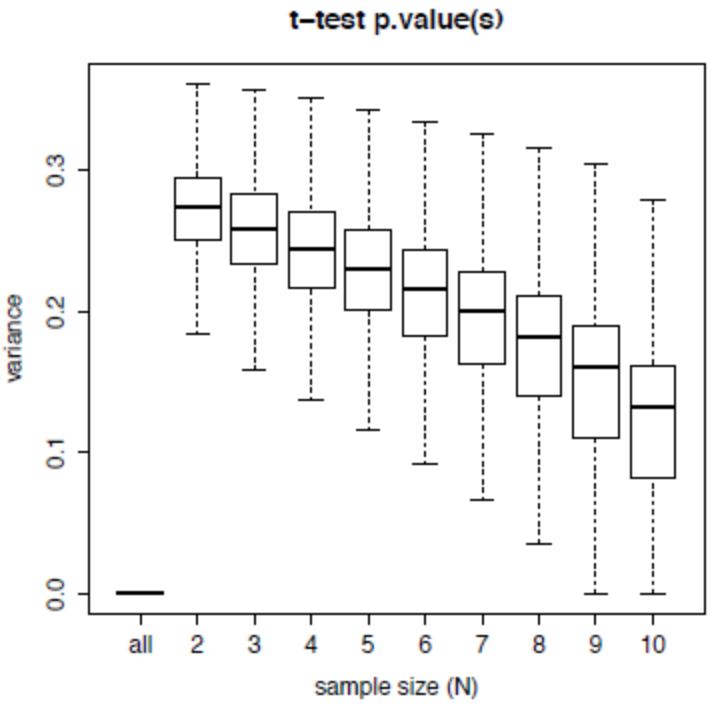
⇒大量的误报



- A biological pathway is a series of actions among molecules in a cell that leads to a certain product or a change in a cell. Thus necessarily the behavour of genes in a pathway is more coordinated than random ones

Issue #2 with ORA

- **It relies on a pre-determined list of DE genes**
- **This list is sensitive to the test statistic used and to the significance threshold used**
- **This list is unstable regardless of the threshold used when sample size is small**



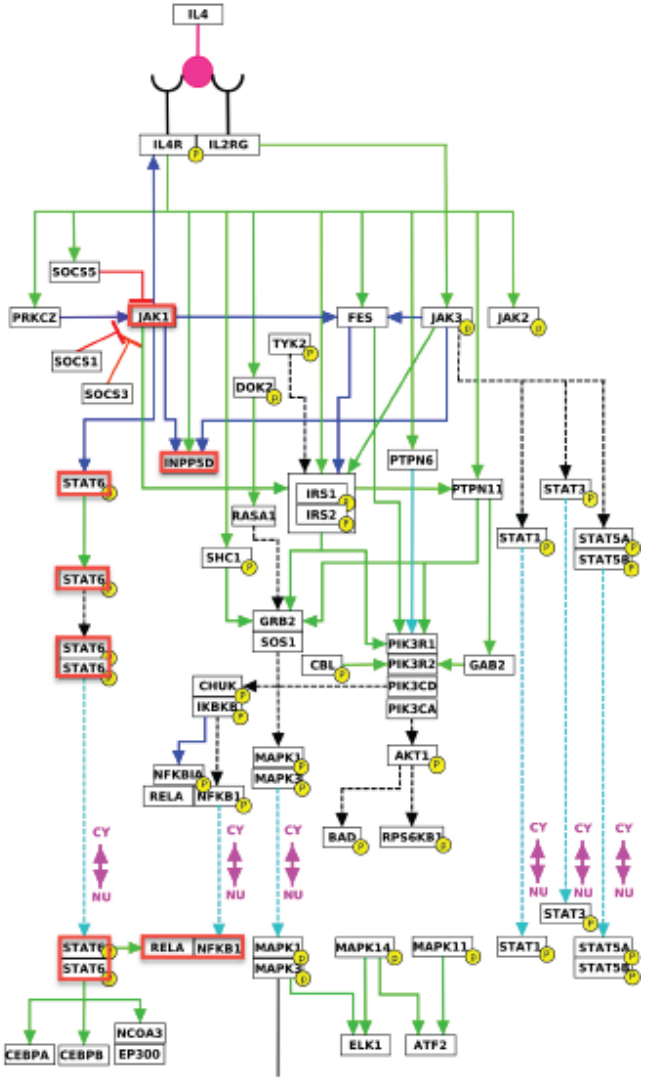
检验统计不够稳定

Issue #3 with ORA

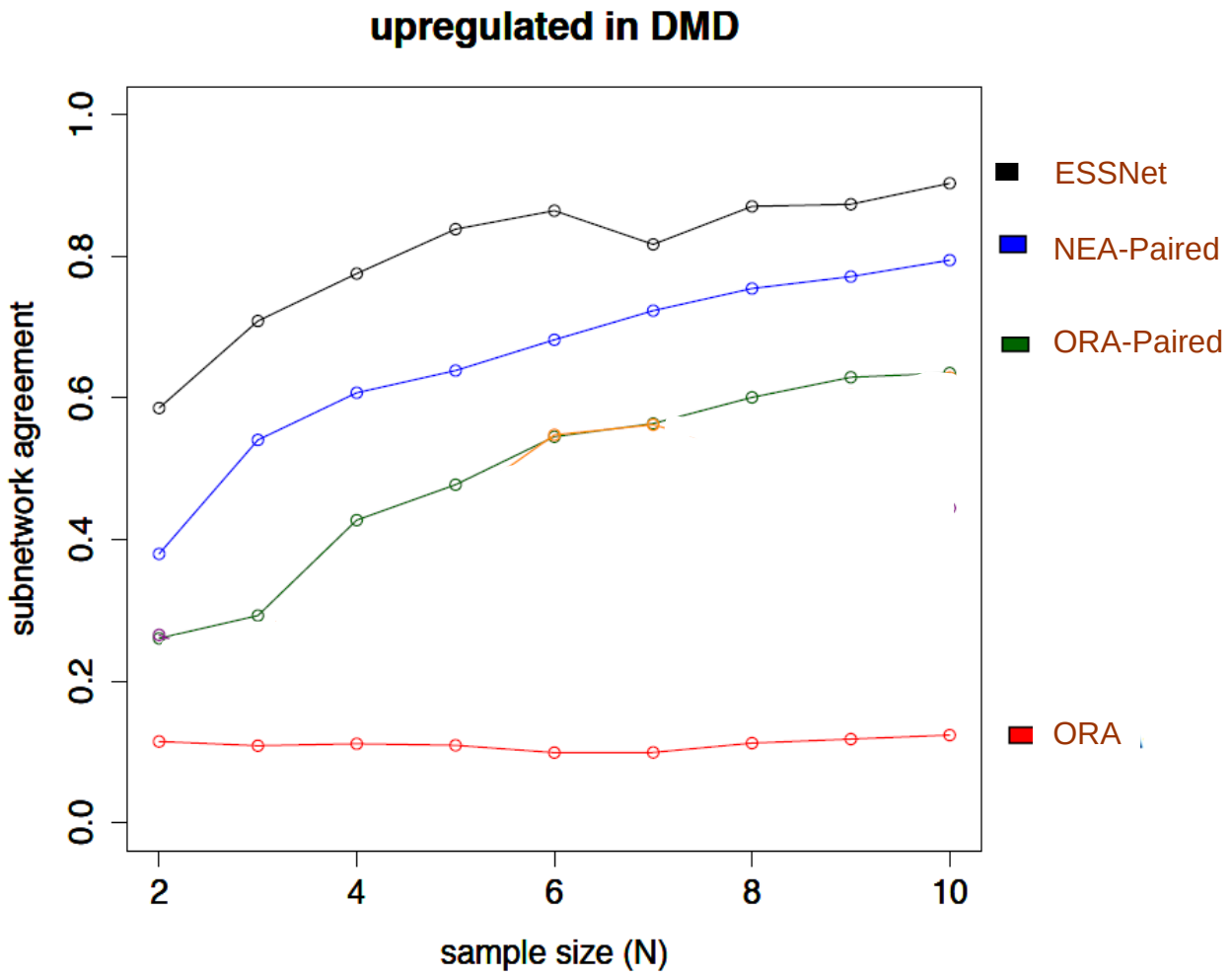
- 它测试是否整个生物途径有显著差异表达

⇒如果生物途径的仅一个分支是与疾病相关，该生物学途径不相关部分的噪声恐会稀释该信号

⇒更多不一定是更好



解决了此三个问题，性能显著提升



总结

- 大数据可以提供一个更全面的了解，填补空白，等等。
- 大数据也将噪声引入分析
- 除非你知道如何驯服这种噪音，更多的数据可能不会导致更好的分析

- **Big data can offer a more complete picture, fill in gaps, etc.**
- **More data can also introduce noise into an analysis**
- **Unless you know how to tame this noise, more data may not lead to a better analysis**